Statistical Methods for Estimating Treatment Effects Based on Boundary Inflated Data

by

Shulin Qing

A thesis submitted to The Johns Hopkins University in conformity with the requirements for the degree of Master of Science.

Baltimore, Maryland

April, 2018

Abstract

This thesis is devoted to developing, evaluating and applying a new statistical method for drawing inference about the effect of a treatment on an outcome that takes values within and on the boundary of a interval of the real line. The methodology is motivated by interest in comparing treatments for employed individuals who experience a traumatic injury with respect to productivity loss.

Primary Reader: Daniel Scharfstein

Secondary Reader: Elizabeth Colantuoni

Acknowledgments

I would like to express my sincere gratitude to my thesis advisor, Dr. Daniel Scharfstein, for introducing me to this research topic, and for his continuous support, enthusiasm and immense knowledge during the course of my thesis. Under his guidance, I have gained a deeper understanding of statistical theory and its applications. I could not have imagined having a better advisor and mentor for my Master's study.

Besides my thesis advisor, I would like to thank my thesis reader, Dr. Elizabeth Colantuoni, for her valuable time and suggestions on this thesis.

I would like to give my special and great appreciation to Dr. Ellen Mackenzie for offering me the opportunity to work on the project of work productivity loss at Major Extremity Trauma Research Consortium (METRC) and for her continuous support and guidance.

My Master's degree could not have been completed without the support from my colleagues and friends in the Biostatistics Department and at METRC. I would like to thank Yanjie Huang, Su Jin Lim, Gege Gui and Yue Cao for helping me solve both academic and non-academic problems and for always encouraging me to pursue my

goals.

Finally, I must express my very profound gratitude to my parents for their considerable help, encouragement and understanding throughout my years of study. This accomplishment would not have been possible without their support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The methods developed in this thesis are motivated by the need to estimate the effect of a non-randomized treatment on an outcome that takes values within and on the boundary of a compact interval of the real line (without loss of generality, consider the interval to be $[0, 1]$). An example of such an outcome is derived from the Work Productivity and Activity Impairment (WPAI) questionnaire, which is a self-administered instrument used to evaluate the impact of disease on productivity. The questionnaire can be used to compute three measures of productivity:

- Absenteeism (A) defined, for employed individuals, as the number of work hours missed due to problems divided by the sum of this number and number of hours actually worked and, for non-employed individuals, to be 1.0.

- Presenteeism (P) defined, for employed individuals, as how much health affected productivity at work on a scale from zero to one;

- Productivity Loss (L) defined as $A + (1 - A)P$.

Figure 1.1 presents the distribution of productivity loss six months after employed patients experienced a traumatic lower limb injury, stratified by treatment (amputation vs. salvage). Notice that the distribution exhibits positive probability mass at 0 and at 1.

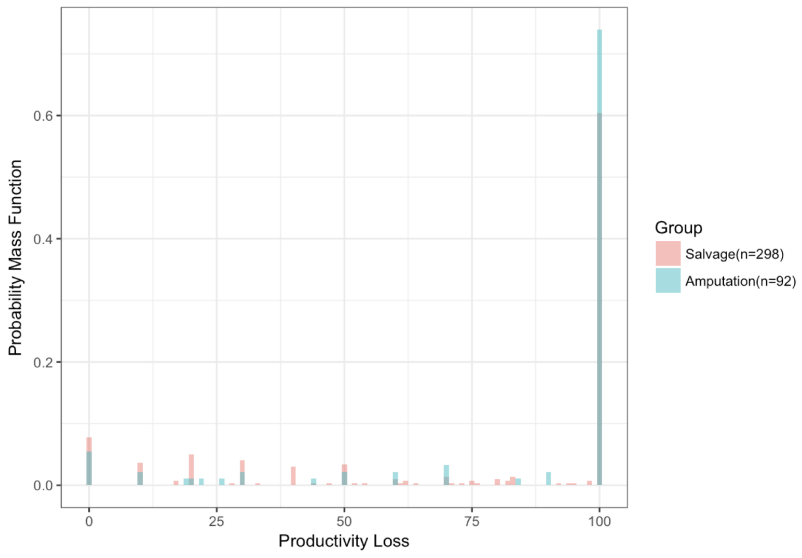

Figure 1.1: The Probability Mass Function of Productivity Loss at 6 Months

A variety of regression methods have be proposed for analyzing boundary inflated data: zero/one inflated Beta regression, latent variable regression, re-scaled outcome Beta regression, and logistic quantile regression.

1.1 Zero/One Beta Regression

Ospina and Ferrari [1] considered the following generative model (GM1) for the outcome Y:

1. Generate $U \sim \text{Bernoulli}(p)$;
2. If $U = 1$, set $Y = 0$ and stop.
3. If $U = 0$, generate $V \sim \text{Bernoulli}(q)$;
4. If $V = 1$, set $Y = 1$ and stop.
5. If $V = 0$, generate $Y \sim \text{Beta}(\mu, \phi)$ and stop.

Notice that q is the conditional probability that $V = 1$ given $U = 0$, which is the same as the conditional probability that $Y = 1$ given $Y \neq 0$. Further, the Beta distribution is parameterized through its mean (μ) and dispersion(ϕ). Under GM1,

$$dF(y) = \begin{cases} p & y = 0 \\ (1-p)q & y = 1 \\ (1-p)(1-q)\frac{y^{\phi\mu-1}(1-y)^{\phi(1-\mu)-1}}{\text{B}(\phi\mu,\phi(1-\mu))} & 0 < y < 1 \end{cases} \tag{1.1}$$

where $\text{B}(\alpha, \beta)$ is the Beta function. They considered an alternative generative model (GM2) of the following form:

1. Generate $U \sim \text{Multinomial}(p', q', 1 - p' - q')$ with corresponding levels 0,1,2;

2. If $U = 0$, set $Y = 0$; If $U = 1$, set $Y = 1$; If $U = 2$, generate $Y \sim \text{Beta}(\mu, \phi)$.

Under GM2,

$$dF(y) = \begin{cases} p' & y = 0 \\ q' & y = 1 \\ (1 - p' - q') \frac{y^{\phi\mu - 1}(1-y)^{\phi(1-\mu)-1}}{\text{B}(\phi\mu, \phi(1-\mu))} & 0 < y < 1 \end{cases} \tag{1.2}$$

Liu and Kong [2] extended GM1 to handle covariates as well as clustered/correlated data through random effects. For purposes of this thesis, it is sufficient to review their extension to handling covariates (X). Specifically, they assume that p, q, μ and ϕ are all functions of X and unknown parameters. They consider the following types of models:

$$\text{logit } \{p(X)\} = X^\top \gamma_p \tag{1.3}$$

$$\text{logit } \{q(X)\} = X^\top \gamma_q \tag{1.4}$$

$$\text{logit } \{\mu(X)\} = X^\top \gamma_\mu \tag{1.5}$$

$$\log\{\phi(X)\} = X^\top \gamma_\phi, \tag{1.6}$$

where γ_p, γ_q, γ_μ and γ_ϕ are parameter vectors of the same dimension as X. Liu and Kong [2] used the Bayesian inferential framework to draw inference about the regression parameters. They implemented their methodology in an R package called **zoib**.

A natural modification of Liu and Kong's GM1 regression approach to GM2 is to replace the logistic regression models for $p(X)$ and $q(X)$ by a multinomial regression

model for $p'(X)$ and $q'(X)$, i.e.,

$$\log\left\{\frac{p'(X)}{1-p'(X)-q'(X)}\right\} = X^\top \gamma_{p'} \tag{1.7}$$

$$\log\left\{\frac{q'(X)}{1-p'(X)-q'(X)}\right\} = X^\top \gamma_{q'} \tag{1.8}$$

This formulation will be relevant for our proposal in Chapter 2.

1.2 Latent Variable Regression

Lesaffre, Rizopoulos and Tsonaka [3] proposed a binomial-logit-normal model for modeling proportions. The basis of their approach is the use of a latent probability U that is assumed to follow a logit-normal distribution, i.e., logit$\{U\}$ is normally distributed.

Let N be the measure-specific denominator, which may vary between individuals. Lesaffre, Rizopoulos and Tsonaka [3] assume that the outcome Y for an individual can be expressed as R/N where,

- the conditional distribution of R given U, N and X is Binomial with sample size N and probability of success U
- the conditional distribution of U given X follows a logit-normal model with mean $X^\top\gamma$ and variance σ^2.

The regression coefficients in this model quantify the effect of covariates on the latent probability U, rather than the outcome Y. This model cannot be employed to analyze

productivity loss data because N is not defined for individuals who are not employed.

1.3 Re-Scaled Outcome Beta Regression

Smithson and Verkuilen [4] suggested re-scaling the outcome and then analyzing the re-scaled outcome using Beta regression. Specifically, they suggested creating a new outcome $Y_{new} = \frac{\epsilon}{2} + (1 - \epsilon)\, Y$ for some small $\epsilon > 0$. This new outcome has a minimum of $\epsilon/2$ and a maximum of $1 - \epsilon/2$. Thus, it is strictly between 0 and 1. The authors argued for its utility because it preserves the mean of the outcome. However, it is not a reasonable for the analysis of work productivity as 0 and 1 have specific substantive meaning.

1.4 Logistic Quantile Regression

Bottai, Cai and McKeown [5] proposed a logistic quantile regression model for bounded continuous outcomes. To deal with outcomes on the boundary, they defined the following ϵ-modified logistic function:

$$h_\epsilon(y) = \log\left(\frac{y + \epsilon}{1 - y + \epsilon}\right), \tag{1.9}$$

where ϵ is some small positive number. Letting $Q_{Y|X}(p)$ be the pth quantile of the conditional distribution of Y given X, they assumed that

$$h_\epsilon(Q_{Y|X}(p)) = \gamma_p^\top X, \tag{1.10}$$

where γ_p are quantile-specific unknown parameters of the same dimension as X. This model implies that

$$Q_{Y|X}(p) = \frac{\exp\{\gamma_p^\top X\}(1+\epsilon) - \epsilon}{1 + \exp\{\gamma_p^\top X\}} \tag{1.11}$$

Here γ_p quantifies the effect of covariates on the pth quantile, although its interpretation is affected by choice of ϵ. By estimating effects for multiple choices of p, one can gain a better understanding of how covariates affect different aspect of the conditional distribution of Y given X. Standard quantile reqression software applied to $h_\epsilon(Y)$ can be used to estimate γ_p [6].

1.5 Discussion

As discussed above, the latent variable regression approach can not be used to analyze productivity loss data. While the remaining approaches can be employed, they are not ideal for obtaining a *parsimonious* measure of the adjusted effect of a treatment on an outcome which takes values within and on the boundary of a compact interval of the real line. To see this, suppose X is partitioned to include an intercept, a treatment indicator (T) and additional covariates (W). The zero/one Beta regression approach would require the reporting of seperate treatment effects in models (1.3) to (1.6). The other two approaches allow for a parsimonious measure of treatment effect. The re-scaled outcome and quantile regression approaches require "ϵ-adjustments" to the outcome and a transformation of the outcome, respectively, in order to employ

standard regression techniques. For these latter approaches, sensitivity of results to choice of ϵ would be necessary.

In this thesis, an alternative approach is presented for estimating the adjusted effect of treatment on a boundary inflated outcome that (1) is parsimonious and (2) does not require "ϵ-adjustment". The approach uses exponential tilting [7] to link the conditional distribution of Y given W and $T = 1$ to the conditional distribution of Y given W and $T = 0$ through a scalar parameter (α) that quantifies the adjusted effect of treatment.

1.6 Outline

In Chapter 2, the core ideas underlying the new approach are presented. In Chapter 3, the new method is applied to the analysis of work productivity of amputees versus salvages. In Chapter 4, a realistic simulation study evaluates the performance of the proposed approach. The final chapter provides concluding remarks.

Chapter 2

Statistical Methods

2.1 Model

Let Y denote the outcome taking values in the interval $[0, 1]$. Let T denote treatment group indicator (1 for treatment, 0 for control). Let W denote a $k \times 1$ covariate vector (including 1 as first entry). Our model is built by (1) modeling the conditional distribution of Y given W among control patients and (2) using exponential tilting to model the conditional distribution of Y given W among treated patients. First, we assume

$$\begin{aligned} \log\left\{\frac{P[Y=0|T=0,W]}{P[0<Y<1|T=0,W]}\right\} &= W'\gamma_1 \\ \log\left\{\frac{P[Y=1|T=0,W]}{P[0<Y<1|T=0,W]}\right\} &= W'\gamma_2, \end{aligned} \tag{2.1}$$

where γ_1 and γ_2 are unknown parameter vectors. Second, we assume that the conditional distribution of Y given $0 < Y < 1$, W and $T = 0$ follows a Beta regression

model. That is,

$$
\begin{aligned}
&dF(y|0<Y<1,W,T=0)\\
&=\frac{\Gamma\{\exp(\phi)\}}{\Gamma\Big(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\Big)\Gamma\Big(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\Big)}y^{\Big(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\Big)-1}(1-y)^{\Big(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\Big)-1},
\end{aligned}
\tag{2.2}
$$

where $0<y<1$ and ϕ and γ_3 are unknown parameters. Together, these assumptions indicate that the conditional distribution of Y given W and $T=0$ has the following form:

$$
dF(y|W,T=0)=\begin{cases}
\frac{\exp(W'\gamma_1)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)} & y=0\\
\frac{1}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)}\frac{\Gamma\{\exp(\phi)\}}{\Gamma\Big(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\Big)\Gamma\Big(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\Big)} & \\
\qquad y^{\Big(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\Big)-1}(1-y)^{\Big(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\Big)-1} & 0<y<1\\
\frac{\exp(W'\gamma_2)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)} & y=1
\end{cases}
\tag{2.3}
$$

The last part of the model specification is to use exponential tilting to connect the conditional distribution of Y given W and $T=1$ to the conditional distribution of Y given W and $T=0$, i.e.,

$$
dF(y|W,T=1)=\frac{dF(y|W,T=0)\exp(\alpha y)}{\int_0^1\exp(\alpha y')dF(y'|W,T=0)}\quad 0\le y\le 1,
\tag{2.4}
$$

where α is an unknown parameter. The integral in the denominator of this expression can be expressed as:

$$\begin{aligned} c(W;\theta) &= \frac{\exp(W'\gamma_1)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)} + \\ &\int_0^1 \frac{\exp(\alpha y)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)} \frac{\Gamma\{\exp(\phi)\}}{\Gamma\left(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\right)\Gamma\left(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\right)} \\ &\times y^{\left(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\right)-1}(1-y)^{\left(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\right)-1}dy + \\ &\frac{\exp(W'\gamma_2)\exp(\alpha)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)}, \end{aligned} \tag{2.5}$$

where $\theta = (\gamma_1', \gamma_2', \gamma_3', \phi, \alpha)'$. Together, these assumptions indicate that the conditional distribution of Y given W and $T = 1$ has the following form:

$$dF(y|W,T=1) = \frac{1}{c(W;\theta)} \begin{cases} \frac{\exp(W'\gamma_1)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)} & y=0 \\ \frac{\exp(\alpha y)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)} \frac{\Gamma\{\exp(\phi)\}}{\Gamma\left(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\right)\Gamma\left(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\right)} & \\ \quad \times y^{\left(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\right)-1}(1-y)^{\left(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\right)-1} & 0<y<1 \\ \frac{\exp(W'\gamma_2)\exp(\alpha)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)} & y=1 \end{cases} \tag{2.6}$$

2.2 Interpretation of α

Under Model (2.6),

$$R(y|W;\theta) = \frac{dF(y|W,T=1)}{dF(y|W,T=0)} = \frac{\exp(\alpha y)}{c(W;\theta)}, \tag{2.7}$$

where $R(y|W;\theta)$ is the ratio of the conditional (on W) distribution of Y at y for treated versus control patients. Thus,

$$\log\left\{\frac{R(y|W;\theta)}{R(y'|W;\theta)}\right\} = \alpha(y - y'). \tag{2.8}$$

Using Bayes' rule, it can be shown that

$$\log\left\{\frac{\text{odds}(P[T=1|W,Y=y])}{\text{odds}(P[T=1|W,Y=y'])}\right\} = \alpha(y - y'). \tag{2.9}$$

The parameter α can be interpreted as either:

- the difference in the log ratio of the conditional (on W) distribution of Y at y for treated versus control patients per unit change in y, or
- the difference in the log conditional (on W and $Y = y$) odds of treatment per unit change in y

Thus, $\alpha > 0$ (< 0) indicates that treated patients are more (less) likely to have higher values of the outcome than control patients, after adjusting for W. Given the familiarity of odds ratios from logistic regression models, it may be easier to use the interpretation of α from (2.9).

2.3 Inference

To estimate θ, we use the method of maximum likelihood. The log likelihood for θ based on a single observation $O = (T, W, Y)$ is

$$\begin{aligned}\ell(O;\theta) &= I(Y=0)\{W'\gamma_1\} + I(Y=1)\{W'\gamma_2\} - \log\{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)\} + \\ &\quad I(0<Y<1)g(W,Y;\theta) + T\left\{\alpha Y - \log c(W;\theta)\right\}, \end{aligned} \tag{2.10}$$

where

$$\begin{aligned}g(W,Y;\theta) &= \log\{\Gamma\{\exp(\phi)\} - \log\left\{\Gamma\left(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\right)\right\} - \\ &\quad \log\left\{\Gamma\left(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\right)\right\} + \left(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\right)\log(Y) + \\ &\quad \left(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\right)\log(1-Y). \end{aligned} \tag{2.11}$$

Note that

$$\begin{aligned}c(W;\theta) &= \frac{\exp(W'\gamma_1)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)} + \\ &\quad \int_0^1 \frac{\exp(\alpha y)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)}\exp(g(W,y;\theta))dy + \\ &\quad \frac{\exp(W'\gamma_2)\exp(\alpha)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)}. \end{aligned} \tag{2.12}$$

Based on a single observation, the score for γ_1 is

$$S_{\gamma_1}(O;\theta) = \left\{I(Y=0) - \frac{\exp(W'\gamma_1)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)}\right\}W - T\frac{\frac{\partial c(W;\theta)}{\partial\gamma_1}}{c(W;\theta)}; \tag{2.13}$$

the score for γ_2 is

$$S_{\gamma_2}(O;\theta) = \left\{I(Y=1) - \frac{\exp(W'\gamma_2)}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)}\right\}W - T\frac{\frac{\partial c(W;\theta)}{\partial\gamma_2}}{c(W;\theta)}; \tag{2.14}$$

the score for γ_3 is

$$S_{\gamma_3}(O;\theta) = I(0 < Y < 1)\frac{\partial g(W,Y;\theta)}{\partial \gamma_3} - T\frac{\frac{\partial c(W;\theta)}{\partial \gamma_3}}{c(W;\theta)}; \tag{2.15}$$

the score for ϕ is

$$S_{\phi}(O;\theta) = I(0 < Y < 1)\frac{\partial g(W,Y;\theta)}{\partial \phi} - T\frac{\frac{\partial c(W;\theta)}{\partial \phi}}{c(W;\theta)}; \tag{2.16}$$

and the score for α is

$$S_{\alpha}(O;\theta) = T\left\{Y - \frac{\frac{\partial c(W;\theta)}{\partial \alpha}}{c(W;\theta)}\right\}, \tag{2.17}$$

where

$$\begin{aligned}\frac{\partial g(W,Y;\theta)}{\partial \gamma_3} &= \left\{\frac{\exp(\phi)\exp(W'\gamma_3)}{\{1+\exp(W'\gamma_3)\}^2}\right\} \times \\ &\left\{\log\left(\frac{Y}{1-Y}\right) - \psi\left(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\right) + \psi\left(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\right)\right\} W\end{aligned} \tag{2.18}$$

$$\begin{aligned}\frac{\partial g(W,Y;\theta)}{\partial \phi} &= \frac{\exp(\phi)\exp(W'\gamma_3)}{\{1+\exp(W'\gamma_3)\}}\left\{\log(Y) - \psi\left(\frac{\exp(\phi)\exp(W'\gamma_3)}{1+\exp(W'\gamma_3)}\right)\right\} + \\ &\frac{\exp(\phi)}{\{1+\exp(W'\gamma_3)\}}\left\{\log(1-Y) - \psi\left(\frac{\exp(\phi)}{1+\exp(W'\gamma_3)}\right)\right\} + \\ &\exp(\phi)\psi\left(\exp(\phi)\right)\end{aligned} \tag{2.19}$$

$$\begin{aligned}\frac{\partial c(W;\theta)}{\partial \gamma_1} &= \left\{\frac{\exp(W'\gamma_1)(1+\exp(W'\gamma_2))}{\{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)\}^2} - \right. \\ &\int_0^1 \frac{\exp(W'\gamma_1)\exp(\alpha y)}{\{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)\}^2}\exp(g(W,y;\theta))dy - \\ &\left.\frac{\exp(W'\gamma_1)\exp(W'\gamma_2)\exp(\alpha)}{\{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)\}^2}\right\} W\end{aligned} \tag{2.20}$$

$$\frac{\partial c(W;\theta)}{\partial \gamma_2} = \left\{ \frac{-\exp(W'\gamma_1)(\exp(W'\gamma_2)}{\{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)\}^2} - \int_0^1 \frac{\exp(W'\gamma_2)\exp(\alpha y)}{\{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)\}^2}\exp(g(W,y;\theta))dy + \frac{(1+\exp(W'\gamma_1))\exp(W'\gamma_2)\exp(\alpha)}{\{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)\}^2}\right\} W \tag{2.21}$$

$$\frac{\partial c(W;\theta)}{\partial \gamma_3} = \int_0^1 \frac{\exp(\alpha y)\frac{\partial g(W,Y;\theta)}{\partial \gamma_3}\exp(g(W,y;\theta))}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)}dy \tag{2.22}$$

$$\frac{\partial c(W;\theta)}{\partial \phi} = \int_0^1 \frac{\exp(\alpha y)\frac{\partial g(W,Y;\theta)}{\partial \phi}\exp(g(W,y;\theta))}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)}dy \tag{2.23}$$

$$\frac{\partial c(W;\theta)}{\partial \alpha} = \frac{\exp(\alpha)\exp(W'\gamma_2) + \int_0^1 y\exp(\alpha y)\exp(g(W,y;\theta))dy}{1+\exp(W'\gamma_1)+\exp(W'\gamma_2)}. \tag{2.24}$$

Now, let $S(O;\theta) = (S_{\gamma_1}(O;\theta)', S_{\gamma_2}(O;\theta)', S_{\gamma_3}(O;\theta)', S_{\phi}(O;\theta)', S_{\alpha}(O;\theta)')$. The MLE for θ, $\widehat{\theta}_n$, solves

$$\sum_{i=1}^{n} S(O_i;\theta) = 0. \tag{2.25}$$

From the theory of maximum likelihood, we know that $\widehat{\theta}_n$ converges in probability to θ and $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in distribution to a multivariate normal with mean zero and variance-covariance matrix $I(\theta)^{-1}$, where $I(\theta) = E[S(O;\theta)S(O;\theta)']$. In finite samples,

$$\widehat{\theta}_n \approx N(\theta, I_n(\theta)^{-1}) \tag{2.26}$$

where $I_n(\theta) = nI(\theta)$. We can estimate $I_n(\theta)$ by $\widehat{I}_n(\widehat{\theta}_n)$ where

$$\widehat{I}_n(\theta) = \sum_{i=1}^{n} S(O_i;\theta)S(O_i;\theta)'. \tag{2.27}$$

The estimated standard errors of the components of θ can be computed by taking the square root of the diagonal elements of $\widehat{I}_n(\widehat{\theta}_n)^{-1}$. A 95% Wald confidence interval for the jth component of θ can be computed as the jth component of $\widehat{\theta}_n$ plus/minus 1.96 times the corresponding estimated standard error.

Chapter 3

Data Analysis

The Military Extremity Trauma Research Consortium (METRC) is funded by the United States Department of Defense to "conduct multi-center clinical research studies relevant to the treatment and outcomes of orthopaedic trauma sustained in the military" [8]. For patients experiencing a traumatic lower limb injury, a key research question is to compare the outcomes of those treated with an amputation versus those treated by limb salvage. There is particular interest in comparing these treatment groups, among those employed prior to injury, with respect work productivity at fixed points in time after injury.

To illustrate the methodology developed in this thesis, data from two METRC studies were pooled:

1. TAOS is a randomized controlled trial comparing trans-tibial amputation with and without tibia-fibula synostosis among patients requiring unilateral ampu-

tation following major limb trauma.

2. OUTLET is a prospective cohort study of patients with severe distal tibia, ankle, and/or foot trauma treated with limb salvage or amputation.

TAOS is an ongoing study; data for this analysis were downloaded as of March 1, 2018. The dataset was restricted to individuals who (1) have complete pre-injury (i.e., baseline) information, (2) were employed pre-injury and (3) had complete information on productivity loss at six months post-injury (Y). With these restrictions, the dataset included 298 patients treated with limb salvage ($T = 0$) and 92 treated with amputation ($T = 1$).

To adjust for baseline characteristics between salvages and amputees, the covariate vector W was defined as four indicator variables, representing membership in the second, third, fourth and fifth quintiles of the distribution of the estimated conditional probability of amputation given the following pre-injury variables: age, gender, education, work demand, number of co-morbidities, primary occupation, martial status, pre-injury health status, injury severity score and health insurance (see Table 3.1 for summary statistics associated with these factors, separately for amputees and salvages). This conditional probability is often referred to as the propensity score [9]; it was estimated using logistic regression. The results of the logistic regression model are displayed in Table 3.2. Males (versus females), patients with more demanding jobs (versus those whose jobs are not demanding) and patients with three or more co-morbidities (versus those with none) are more likely to be amputated. Table 3.3

shows the treatment-specific number of patients who fall into each of the quintiles. Figure 3.1 shows the distribution of the propensity scores, stratified by treatment. These results show that there is overlap in the distribution of the propensity scores between the two treatment groups.

3.1 New Approach

Table 3.4 displays the results of fitting the proposed model without covariates. Here, the estimated treatment effect is 0.79 (95% CI: 0.07 to 1.51). That is, the odds of amputation is estimated to be 2.2 times higher for patients with complete productivity loss than patients with no productivity loss. The top panel of Figure 3.2 presents the empirical probability mass functions of productivity loss for amputees and salvages. The bottom panel presents the empirical probability mass function of productivity loss for salvages and the estimated probability mass function of productivity loss for amputeees under the unadjusted exponential tilting model. As can be seen, the figures are very similar showing that the unadjusted exponential tilting model provides a reasonable fit to the observed data.

Table 3.5 shows the results of fitting the proposed model with the propensity score quintiles as covariates. After adjustment, the estimated treatment effect is attenuated to 0.45 (95% CI: -0.40 to 1.29). That is, the adjusted odds of amputation is estimated to be 1.6 times higher for patients with complete productivity loss than patients

with no productivity loss. After adjustment, amputees still tend to have higher productivity loss than salvages but the evidence is substantially weaker. Similar figures (not shown) to Figure 3.2 constructed for each quintile shows that adjusted exponential tilting model provides a reasonable fit to the observed data.

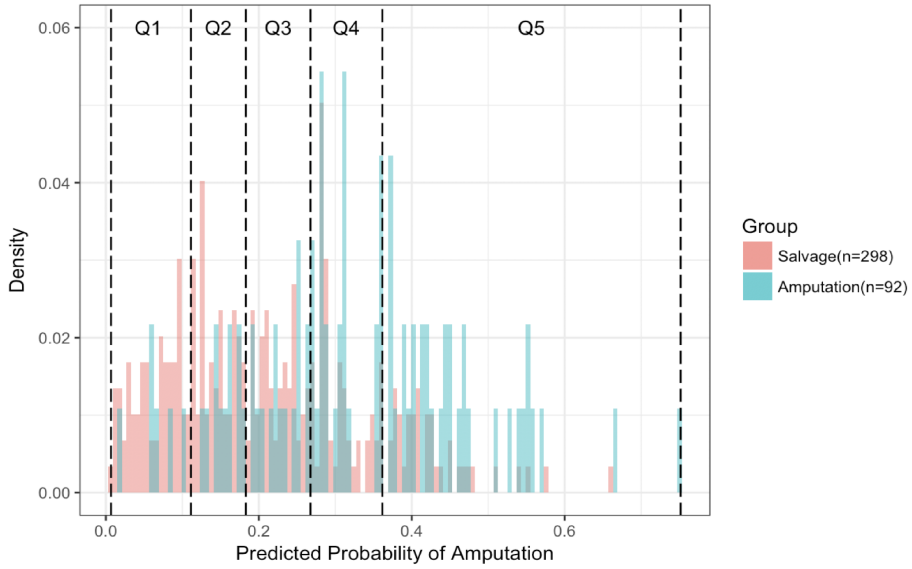

Figure 3.1: The Propensity Score Distribution

Characteristics	**Salvages (n = 298)**	**Amputees (n = 92)**
Age, mean (SD)	38.85 (11.96)	40.46 (11.54)
Gender, No. (%)		
Female	99 (33)	14 (15)
Male	199 (67)	78 (85)
Education, No. (%)		
Less than high school	40 (13)	14 (15)
High school or GED	97 (33)	43 (47)
Some college	161 (54)	35 (38)
Work Demand, No. (%)		
Not at all demanding	27 (9)	1 (1)
Not too demanding	30 (10)	8 (9)
Somewhat demanding	94 (32)	29 (31)
Very demanding	147 (49)	54 (59)
Comorbidity, No. (%)		
0	184 (62)	52 (57)
1-2	107 (36)	34 (37)
3+	7 (2)	6 (6)
Primary Occupation, No. (%)		
Active duty	4 (1)	2 (2)
Working (<35 hrs per week)	31 (11)	10 (11)
Working (≥35 hrs per week)	263 (88)	80 (87)
Marital Status, No. (%)		
Married or cohabiting	143 (48)	48 (52)
Never married	104 (35)	22 (24)
Widowed, divorced, sep	51 (17)	22 (24)
Pre-injury Health, No. (%)		
Fair/Poor	18 (6)	5 (5)
Good	77 (26)	19 (21)
Excellent/Very Good	203 (68)	68 (74)
Injury Severity Score, No. (%)		
0-8	197 (66)	66 (71)
9-12	24 (8)	8 (9)
13-17	36 (12)	10 (11)
18+	41 (14)	8 (9)
Health Insurance, No. (%)		
None	56 (19)	14 (15)
Medicaid	20 (7)	6 (7)
Other insurance	62 (21)	24 (26)
Private	160 (53)	48 (52)

Table 3.1: Characteristics of Amputees and Salvages

	Estimate	Std. Error	95% CI
Intercept	-4.31	1.66	[-7.56,-1.07]
Age	0.00	0.01	[-0.03,0.03]
Male (Ref: Female)	1.13	0.35	[0.44,1.82]
Education (Ref: Less than high school)			
High School or GED	0.14	0.40	[-0.65,0.93]
Some College	-0.47	0.42	[-1.29,0.35]
Work Demand (Ref: Not at all demanding)			
Not too demanding	1.88	1.11	[-0.3,4.05]
Somewhat demanding	1.93	1.06	[-0.14,4.01]
Very demanding	2.12	1.05	[0.06,4.18]
Comorbidity (Ref: 0)			
1-2	0.11	0.29	[-0.45,0.67]
3+	1.76	0.69	[0.41,3.12]
Primary Occupation (Ref: Active duty)			
Working (< 35 hrs per week)	0.16	1.02	[-1.84,2.17]
Working (≥35 hrs per week)	-0.35	0.93	[-2.18,1.48]
Marital Status (Ref: Married or cohabiting)			
Never married	-0.61	0.34	[-1.27,0.06]
Widowed, divorced, sep	0.37	0.34	[-0.31,1.04]
Pre-injury Health (Ref: Fair/Poor)			
Good	0.03	0.62	[-1.19,1.25]
Excellent/Very Good	0.47	0.59	[-0.69,1.63]
Injury Severity Score (Ref: 0-8)			
9-12	0.11	0.49	[-0.85,1.06]
13-17	0.11	0.42	[-0.72,0.94]
18+	-0.04	0.45	[-0.91,0.84]
Health Insurance (Ref: None)			
Medicaid	0.32	0.60	[-0.86,1.49]
Other insurance	0.49	0.45	[-0.38,1.37]
Private	0.47	0.41	[-0.34,1.28]

Table 3.2: Results of Logistic Regression

	Q1	Q2	Q3	Q4	Q5	Total
Amputation	6	11	16	23	36	92
Salvage	72	67	62	55	42	298

Table 3.3: Treatment-Specific Number of Patients in Propensity Score Quintiles

	Estimate	Std. Error	95% CI
γ_1	-1.34	0.21	[-1.76,-0.92]
γ_2	0.68	0.12	[0.45,0.92]
γ_3	-0.19	0.10	[-0.39,0.01]
ϕ	1.10	0.14	[0.83,1.37]
α	0.79	0.37	[0.07,1.51]

Table 3.4: Unadjusted Model

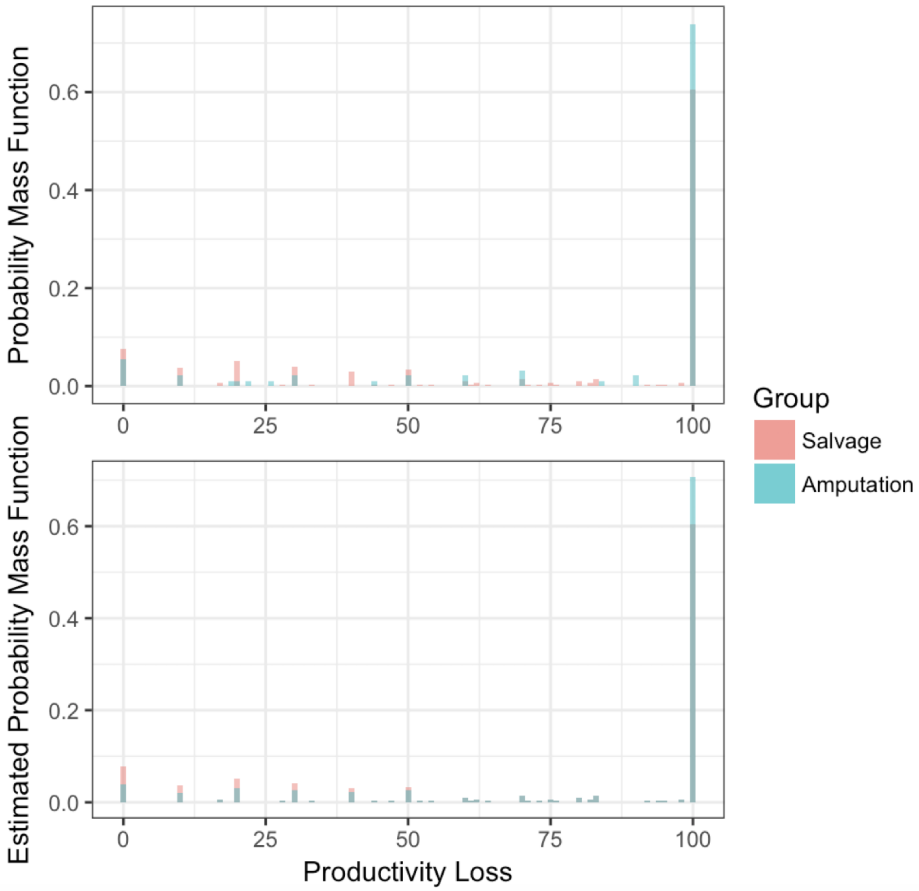

Figure 3.2: Top Panel: Empirical Probability Mass Functions for Amputees and Salvages; Bottom Panel: Empirical Probability Mass Function for Salvages and Estimated Probability Mass Function for Amputees under the Unadjusted Exponential Tilting Model.

	Estimate	**Std. Error**	**95% CI**
$\boldsymbol{\gamma_1}$			
Q1	-1.05	0.37	[-1.77,-0.33]
Q2	-0.18	0.59	[-1.34,0.98]
Q3	-0.05	0.55	[-1.13,1.03]
Q4	-1.20	0.87	[-2.91,0.51]
Q5	-1.11	0.83	[-2.74,0.53]
$\boldsymbol{\gamma_2}$			
Q1	0.28	0.25	[-0.21,0.76]
Q2	0.58	0.36	[-0.13,1.28]
Q3	0.27	0.35	[-0.43,0.96]
Q4	0.68	0.36	[-0.03,1.4]
Q5	0.73	0.37	[0.01,1.45]
$\boldsymbol{\gamma_3}$			
Q1	-0.34	0.21	[-0.76,0.08]
Q2	0.11	0.31	[-0.50,0.71]
Q3	0.31	0.28	[-0.23,0.85]
Q4	0.11	0.36	[-0.60,0.82]
Q5	0.34	0.30	[-0.26,0.93]
$\boldsymbol{\phi}$	1.12	0.16	[0.80,1.43]
$\boldsymbol{\alpha}$	0.45	0.43	[-0.40,1.29]

Table 3.5: Adjusted Model

3.2 Other Approaches

In this section, the application of the alternative regression approaches discussed in Chapter 1 is considered.

3.2.1 Zero/One Beta Regression

Table 3.6 presents the results produced by the R package **zoib**. The results indicate that (1) there is little difference between amputees and salvages with regards to the probability of no productivity loss (see Treatment under $p(X)$), (2) amputees have a higher conditional (on having some productivity loss) probability of full productivity loss (see Treatment under $q(X)$), (3) there is little difference between amputees and salvages with respect the conditional (on have some but not complete productivity loss) mean productivity loss (see Treatment under $\mu(X)$). This is a complex way of presenting the effect of amputation on productivity loss.

3.2.2 Re-Scaled Outcome Beta Regression

Table 3.7 presents the results produced by the R package **betareg**, using three different values of ϵ. The results indicate that amputees have a slightly higher adjusted mean productivity loss than salvages, although no difference cannot be ruled out. The estimated treatment effect appears insensitive to the choice of ϵ.

3.2.3 Logistic Quantile Regression

Table 3.8 presents the results produced by the R package **lqr** with $\epsilon = 0.0001$, for four quantiles. The results show that there is no adjusted differences between the 75th percentile and between the median of amputees and salvages. After adjustment, the 25th percentile for amputees is higher than salvages and the 10th percentile for amputees is lower than salvages, although no difference cannot be ruled out. Tables 3.9 to 3.11 show results for three other choices of ϵ, $\epsilon = 0.001$, $\epsilon = 0.01$ and $\epsilon = 0.1$. The inferences are similar as discussed above.

	Posterior Mean	Posterior Credible Interval 2.5%	97.5%
$\mu(X)$			
Intercept	-0.323	-0.683	0.046
Treatment	-0.014	-0.551	0.534
Q2	0.106	-0.460	0.656
Q3	0.317	-0.233	0.855
Q4	0.123	-0.462	0.704
Q5	0.374	-0.251	0.991
$p(X)$			
Intercept	-1.970	-2.711	-1.335
Treatment	0.003	-1.178	1.006
Q2	-0.612	-1.756	0.427
Q3	-0.277	-1.341	0.732
Q4	-1.958	-3.858	-0.407
Q5	-1.970	-3.985	-0.372
$q(X)$			
Intercept	0.262	-0.220	0.773
Treatment	0.542	-0.048	1.173
Q2	0.572	-0.114	1.270
Q3	0.215	-0.480	0.917
Q4	0.650	-0.071	1.375
Q5	0.623	-0.121	1.370
$\log\{\phi(X)\}$	1.082	0.852	1.311

Table 3.6: Posterior Estimates of Parameters from Zero/One Beta Regression; $\mu(X)$, $p(X)$, $q(X)$ and $\phi(X)$ are defined in Equations (1.3) to (1.6).

	$\epsilon = 0.1$		$\epsilon = 0.01$		$\epsilon = 0.001$	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Intercept	0.51	[0.25,0.77]	0.60	[0.31, 0.89]	0.64	[0.35,0.94]
Treatment	0.14	[-0.14,0.42]	0.13	[-0.17 ,0.43]	0.13	[-0.18 ,0.43]
Q2	0.33	[-0.03,0.68]	0.32	[-0.07,0.72]	0.32	[-0.09,0.72]
Q3	0.20	[-0.16,0.56]	0.19	[-0.21 ,0.58]	0.18	[-0.23 ,0.58]
Q4	0.47	[0.10, 0.83]	0.46	[0.06 ,0.86]	0.45	[0.05 ,0.86]
Q5	0.48	[0.11,0.86]	0.47	[0.06 ,0.88]	0.46	[0.04 ,0.87]
$\log\{\phi\}$	0.54	[0.41,0.66]	-0.17	[-0.30 , -0.05]	-0.56	[-0.68 ,-0.43]

Table 3.7: Parameter Estimates from Re-Scaled Outcome Beta Regression for Different Values of ϵ

	Quantile = 0.1		Quantile = 0.25		Quantile = 0.5		Quantile = 0.75	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Intercept	-7.89	[-10.36,-5.6]	-2.09	[-4.18,-0.01]	3.24	[1.43,5.05]	9.21	[7.13,11.29]
Treatment	-1.59	[-4.01,0.83]	0.83	[-2.08,3.74]	0	[-1.04,1.04]	0	[-2.82,2.82]
Q2	1.05	[-1.92,4.01]	2.53	[-0.6,5.66]	5.97	[3.86,8.08]	0	[-2.90,2.90]
Q3	0.28	[-2.87,3.43]	1.46	[-1.5,4.43]	5.97	[2.77,9.17]	0	[-2.63,2.63]
Q4	5.08	[1.26,8.9]	3.54	[-0.09,7.17]	5.97	[3.31,8.63]	0	[-3.18,3.18]
Q5	4.70	[1.66,7.74]	3.70	[0.17,7.24]	5.97	[3.85,8.09]	0	[-3.15,3.15]

Table 3.8: Estimated Coefficients and Confidence Intervals of Adjusted Logistic Quantile Regression at Four Quantiles ($\epsilon = 0.0001$)

	Quantile = 0.1		Quantile = 0.25		Quantile = 0.5		Quantile = 0.75	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Intercept	-5.99	[-7.78,-4.20]	-2.32	[-3.98,-0.66]	2.38	[1.01,3.76]	6.91	[5.48,8.33]
Treatment	-1.19	[-3.01,0.63]	0.07	[-2.02,2.15]	0	[-2.33,2.33]	0	[-2.04,2.04]
Q2	0.79	[-1.45,3.02]	1.63	[-0.62,3.89]	4.52	[1.83,7.22]	0	[-2.1,2.1]
Q3	0.21	[-2.15,2.58]	0.90	[-1.34,3.14]	4.52	[2.31,6.73]	0	[-1.84,1.84]
Q4	3.80	[0.93,6.67]	3.11	[0.26,5.96]	4.52	[1.59,7.46]	0	[-2.29,2.29]
Q5	3.53	[1.24,5.81]	3.22	[0.63,5.82]	4.52	[1.55,7.50]	0	[-2.19,2.19]

Table 3.9: Estimated Coefficients and Confidence Intervals of Adjusted Logistic Quantile Regression at Four Quantiles ($\epsilon = 0.001$)

	Quantile = 0.1		Quantile = 0.25		Quantile = 0.5		Quantile = 0.75	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Intercept	-4.01	[-5.22,-2.8]	-1.67	[-2.84,-0.50]	1.56	[0.63,2.49]	4.62	[3.66,5.57]
Treatment	-0.66	[-1.83,0.50]	0.08	[-1.34,1.50]	0	[-1.49,1.49]	0	[-1.27,1.27]
Q2	0.51	[-1.01,2.02]	1.03	[-0.55,2.61]	3.05	[1.40,4.71]	0	[-1.37,1.37]
Q3	0.15	[-1.45,1.74]	0.56	[-1.03,2.14]	3.05	[1.53,4.58]	0	[-1.14,1.14]
Q4	2.30	[0.43,4.17]	2.06	[0.13,3.99]	3.05	[1.21,4.90]	0	[-1.50,1.50]
Q5	2.30	[0.76,3.84]	2.11	[0.33,3.89]	3.05	[1.23,4.87]	0	[-1.29,1.29]

Table 3.10: Estimated Coefficients and Confidence Intervals of Adjusted Logistic Quantile Regression at Four Quantiles ($\epsilon = 0.01$)

	Quantile = 0.1		Quantile = 0.25		Quantile = 0.5		Quantile = 0.75	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Intercept	-2.09	[-2.75,-1.44]	-1.00	[-1.67,-0.33]	1.25	[0.75,1.75]	2.40	[1.94,2.85]
Treatment	-0.26	[-0.86,0.33]	0.08	[-0.69,0.84]	0	[-0.78,0.78]	0	[-0.51,0.51]
Q2	0.26	[-0.56,1.07]	0.52	[-0.38,1.43]	1.15	[0.28,2.02]	0	[-0.62,0.62]
Q3	0.08	[-0.78,0.94]	0.27	[-0.64,1.19]	1.15	[0.37,1.93]	0	[-0.49,0.49]
Q4	0.93	[-0.02,1.89]	1.08	[0.02,2.13]	1.15	[0.23,2.06]	0	[-0.73,0.73]
Q5	1.06	[0.24,1.89]	1.07	[0.08,2.06]	1.15	[0.23,2.06]	0	[-0.5,0.5]

Table 3.11: Estimated Coefficients and Confidence Intervals of Adjusted Logistic Quantile Regression at Four Quantiles ($\epsilon = 0.1$)

Chapter 4

Simulation Study

A simulation study was conducted to evaluate the performance of the estimation procedure discussed in Chapter 2. Data for treatment and quintile membership were drawn jointly from a 10-dimensional multinomial distribution with cell probabilities proportional to the counts in Table 3.3. The outcome was then drawn under our proposed model with parameters set to estimates in Table 3.5. In drawing the outcome, rejection sampling was employed [10]. For sample sizes $n = 500$ and $n = 1000$, 1000 datasets were simulated.

Tables 4.1 and 4.2 display the results for $n = 500$ and $n = 1000$, respectively. At sample $n = 500$, there is low bias for all parameters except for the $\boldsymbol{\gamma_1}$ coefficients associated with the fourth and fifth quantiles. Coverage rates of 95% confidence intervals are adequate for all parameters. When the sample size is increased to $n = 1000$, the bias is decreased for all parameter values, but there continues to be higher

bias for the $\boldsymbol{\gamma_1}$ coefficients associated with the fourth and fifth quintiles. Coverage rates for 95% confidence intervals are adequate for all parameters.

	Truth	Mean	Bias	MSE	Coverage
$\boldsymbol{\gamma_1}$					
Q1	-1.053	-1.076	-0.023	0.111	0.954
Q2	-0.177	-0.201	-0.024	0.337	0.961
Q3	-0.048	-0.047	0.001	0.256	0.957
Q4	-1.201	-1.724	-0.523	4.992	0.961
Q5	-1.108	-1.658	-0.550	4.988	0.972
$\boldsymbol{\gamma_2}$					
Q1	0.276	0.277	0.001	0.049	0.957
Q2	0.578	0.608	0.030	0.106	0.941
Q3	0.266	0.270	0.004	0.101	0.944
Q4	0.684	0.685	0.001	0.109	0.943
Q5	0.725	0.729	0.004	0.106	0.951
$\boldsymbol{\gamma_3}$					
Q1	-0.336	-0.322	0.014	0.027	0.953
Q2	0.108	0.088	-0.020	0.066	0.949
Q3	0.308	0.293	-0.015	0.057	0.951
Q4	0.109	0.090	-0.019	0.069	0.943
Q5	0.335	0.321	-0.014	0.066	0.951
$\boldsymbol{\phi}$	1.115	1.160	0.045	0.014	0.944
$\boldsymbol{\alpha}$	0.445	0.490	0.045	0.140	0.963

Table 4.1: Results of 1000 Simulations (Sample Size = 500)

	Truth	Mean	Bias	MSE	Coverage
$\boldsymbol{\gamma_1}$					
Q1	-1.053	-1.070	-0.017	0.055	0.951
Q2	-0.177	-0.172	0.005	0.153	0.945
Q3	-0.048	-0.039	0.009	0.120	0.956
Q4	-1.201	-1.272	-0.071	0.337	0.961
Q5	-1.108	-1.190	-0.082	0.331	0.966
$\boldsymbol{\gamma_2}$					
Q1	0.276	0.278	0.002	0.025	0.953
Q2	0.578	0.595	0.017	0.052	0.948
Q3	0.266	0.270	0.004	0.050	0.950
Q4	0.684	0.677	-0.007	0.051	0.948
Q5	0.725	0.725	0.000	0.053	0.953
$\boldsymbol{\gamma_3}$					
Q1	-0.336	-0.329	0.007	0.013	0.947
Q2	0.108	0.099	-0.009	0.033	0.943
Q3	0.308	0.298	-0.010	0.029	0.950
Q4	0.109	0.100	-0.009	0.033	0.946
Q5	0.335	0.334	-0.001	0.033	0.950
$\boldsymbol{\phi}$	1.115	1.135	0.020	0.006	0.952
$\boldsymbol{\alpha}$	0.445	0.466	0.021	0.066	0.956

Table 4.2: Results of 1000 Simulations (Sample Size = 1000)

Chapter 5

Conclusion

This thesis developed a new statistical method for evaluating the effect of a treatment on an outcome that takes values within and on the boundary of a interval of the real line. The methodology was used to compare amputation versus limb salvage for employed individuals who experienced a traumatic injury with respect to productivity loss. A realistic simulation study was conducted to evaluated the performance of the statistical procedure under correct model specification.

The model in this thesis is highly parametric. It would be interesting to consider models that rely on less parametric assumptions. It would also be useful to extend the methods in this thesis to handle repeated measurements and clustering.

Bibliography

[1] R. Ospina and S. Ferrari, "Inflated Beta Distributions," *Statistical Papers*, vol. 51, pp. 1–15, 06 2007.

[2] F. Liu and Y. Kong, "zoib: An R Package for Bayesian Inference for Beta Regression and Zero/One Inflated Beta Regression," *The R Journal*, vol. 7, no. 2, pp. 34–51, 2015. [Online]. Available: https://journal.r-project.org/archive/2015/RJ-2015-019/index.html

[3] E. Lesaffre, D. Rizopoulos, and R. Tsonaka, "The Logistic Transform for Bounded Outcome Scores," *Biostatistics*, vol. 8, no. 1, pp. 72–85, 2007.

[4] J. Verkuilen and M. Smithson, "Mixed and Mixture Regression Models for Continuous Bounded Responses Using the Beta Distribution," *Journal of Educational and Behavioral Statistics*, vol. 37, no. 1, pp. 82–113, 02 2012.

[5] M. Bottai, B. Cai, and R. E. McKeown, "Logistic Quantile Regression for Bounded Outcomes," *Statistics in Medicine*, vol. 29, pp. 309–317, 2010.

BIBLIOGRAPHY

[6] C. E. Galarza, L. Benites, and V. H. Lachos, "lqr: Robust Linear Quantile Regression," 2018, r package version 1.7. [Online]. Available: https://CRAN.R-project.org/package=lqr

[7] O. E. Barndorff-Nielsen and D. R. Cox, "Inference and Asymptotics," *London: Chapman Hall*, 1994.

[8] E. J. MacKenzie, "Building a Clinical Research Network in Trauma Orthopaedics: The Major Extremity Trauma Research Consortium (METRC)," *Journal of Orthopaedic Trauma*, vol. 30, pp. 353–361, 2016.

[9] P. R. Rosenbaum and D. B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, vol. 70, pp. 41–55, 1983.

[10] C. P. Robert and G. Casella, "Monte Carlo Statistical Methods," *Springer New York*, pp. 47–53, 2004.

Vita

Shulin Qing received the Bachelor of Science degree with Honor in Biochemistry and Cell Biology from University of California, San Diego in June, 2016, and enrolled in the Biostatistics Master of Science program at Johns Hopkins University in September, 2016. She was inducted into the UCSD Muir College Caledonian Society and Phi Beta Kappa Honor Society in 2016.

Starting in July 2017, Shulin has been working at METRC and her research mainly focuses on investigating work productivity loss in patients with major extremity trauma.